MEGAnnotator

Multi-threaded Enhanced prokaryotic Genome Annotator

Lugli Gabriele Andrea

November 23, 2015

For any suggestion or problem related to MEGAnnotator: gabrieleandrea.lugli@studenti.unipr.it

Table of contents

1. What's MEGAnnotator?

MEGAnnotator is a **M**ulti-threaded **E**nhanced prokaryotic **G**enome **A**nnotator. This pipeline allows the generation of an annotated GenBank file fulfilling the NCBI guidelines for assembled microbial genomes submission, based on DNA shotgun sequencing reads, and minimizes manual intervention, removes waiting times between software program executions, while also improving the final quality of both assembly and annotation outputs.

2. What could MEGAnnotator do?

MEGAnnotator has three program sections:

a. Genomic Assembly

Starting from genomic raw reads, MEGAnnotator performs the assembly followed by contigs selection, quality controls, ORFs prediction and genes annotation concluding with the generation of a GenBank file.

b. Metagenomic Assembly

Starting from metagenomics raw reads, MEGAnnotator performs the assembly followed by the ORFs prediction and genes annotation of the resulting contigs.

c. Genes Annotation only

Starting from a pre-assembled genome, MEGAnnotator performs the ORFs prediction and the genes annotation.

3. System requirements

MEGAnnotator should run on all Unix platforms, although it has not tested in all platforms. If you prefer to not install the program in your own system, three versions of MEGAnnotator are provided in VirtualBox with all the dependencies preinstalled and the datasets formatted.

MEGAnnotator-complete: http://probiogenomics.unipr.it/sw/MEGAnnotator-complete.zip including all the software and all the mentioned databases (Pfam-A, NCBI nr and RefSeq) (decompressed size 338GB).

MEGAnnotator-essential: http://probiogenomics.unipr.it/sw/MEGAnnotator-essential.zip including all the software, Pfam-A and NCBI RefSeq databases (decompressed size 109GB).

MEGAnnotator-partial: http://probiogenomics.unipr.it/sw/MEGAnnotator-partial.zip including all the software and Pfam-A database (decompressed size 10GB).

If you choose to download a MEGAnnotator VirtualBox, you can skip the chapter 4, 5 and 6, because you do not need anything else to run the program properly, except for the partial version that need the installation of at least one NCBI database.

4. Installation

First, place the distribution tarball to your work directory. Then, uncompress the distribution tarball and make the files executable typing:

```
unzip MEGAnnotator-master.zip
chmod 755 -R MEGAnnotator-master
```

MEGAnnotator is a bash script, so it is unnecessary to compile. However, to do a complete analysis, several extra programs are invoked by MEGAnnotator. Therefore, before running MEGAnnotator, users should install the programs listed in the next paragraph.

5. Software requirements and dependencies

MEGAnnotator requires the following programs or package for full functionality:

- Java version 1.7 or superior (type "**sudo apt-get install default-jre**" to install)
- readseq (type "**sudo apt-get install readseq**" to install)
- bwa (type "**sudo apt-get install bwa**" to install)
- samtools (type "**sudo apt-get install samtools**" to install)
- tabix (type "**sudo apt-get install tabix**" to install)
- gawk (type "**sudo apt-get install gawk**" to install)
- hmmscan (type "**sudo apt-get install hmmer**" to install)
- emboss software suit (type "**sudo apt-get install emboss**" to install)
- ABySS (downloadable from GitHub https://github.com/bcgsc/abyss)
  Must be included in the PATH.
- RNAmmer (visit http://www.cbs.dtu.dk/services/RNAmmer/ to install)
  Must be included in the PATH.
- tRNAscan-SE (visit http://selab.janelia.org/tRNAscan-SE/ to install)
  Must be included in the PATH.
- GATK (visit https://www.broadinstitute.org/gatk/download/ to install)
  Must be placed in the bin folder of MEGAnnotator. Simply copy the jar file and rename it as "GenomeAnalysisTK.jar" (if different).

6. Databases

To perform the genes annotation is essential to have available the Pfam-A and NCBI (nr or RefSeq) databases. While the Pfam database is pre-formatted available online, the letter need to be formatted using prerapsearch. User database folder path will be requested by MEGAnnotator at each run.

a. Pfam-A database

Visit the Pfam website ftp for the Pfam-A.hmm.gz download:

ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release

Place the compressed file to your own databases folder and decompress the database:

```
gzip -d Pfam-A.hmm.gz
hmmpress Pfam-A.hmm
```

b. NCBI (nr) database

Visit the NCBI website ftp for the nr.gz file download:

ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/

Place the compressed file to your own database folder and decompress the sequences:

```
gzip -d nr.gz
```

From console, move to the MEGAnnotator main directory to build the database with:

```
bin/./prerapsearch -f T -d /folder/nr -n /folder/rapsearch_nr
```

N.B. place the complete path of the decompressed nr file location in the above command (where *folder* is displayed). The amount of disk space needed for the database building is about 200 Gigabyte. It will require several hours.

c. NCBI (bacteria) RefSeq database

To download the last release of the bacteria RefSeq database we suggest executing the following command from console in a specific folder:

```
wget ftp://ftp.ncbi.nlm.nih.gov/refseq/release/bacteria/*.faa.gz
gzip -d *.gz
cat *.faa > bacteria_refseq.fasta
rm *.faa
```

From console, move to the MEGAnnotator main directory to build the database with:

```
bin/./prerapsearch -f T -d /folder/bacteria_refseq.fasta -n /folder/rapsearch_refseq
```

N.B. place the complete path of the RefSeq fasta file location in the above command (where *folder* is displayed). The amount of disk space needed for the database building is about 100 Gigabyte. It will require few hours.

7. Input data

a. Raw Data

Raw data should be supplied as fastq file, furthermore MEGAnnotator is capable to manage paired-end illumina data.

b. Reference Genome (optional)

Reference Genome should supplied as fasta file. The reference genome is optional; its usage is limited in case users would order the obtained contigs. However, the assembly program does not use the reference genome to perform the contigs creation.

N.B. for a demonstration, whole genome sequencing test data with the corresponding reference genome are downloadable at:

http://probiogenomics.unipr.it/sw/MEGAnnotator-dataset.zip

These test data are included within every VirtualBox provided.

8. Output data

a. Assembly results

Within the assembly output, you can found the multifasta file containing the contigs, the info file with the assembly output information and the assembly log file.

b. Alignment results (optional)

Output file regarding the final alignment performed against the reference genome.

c. Improvement quality results

Multifasta file containing the improved contigs, tabular file with the nucleotidic substitutions and vcf files.

d. Annotation results

The result consist in a GenBank file. Within the GenBank file, the contigs are represented by fasta_records. For a correct visualization of the GenBank file, we suggest the utilization of the free software Artemis (https://www.sanger.ac.uk/resources/software/artemis/). Furthermore, the annotated formats GFF3, EMBL and XML were also provided.

9. Usage

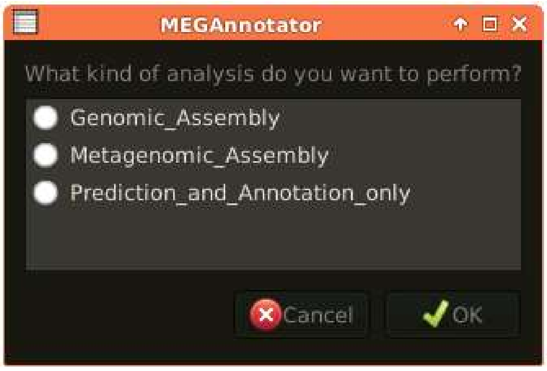
To correctly run MEGAnnotator.sh, the bin and lib folders must be located in the MEGAnnotator script directory, as well as the script annotation.sh, genomic.sh and metagenomic.sh. Please do not change any file within these directory, otherwise the pipeline may be compromised.

Simply run the script typing:

```
./MEGAnnotator.sh
```

10. Tutorial

MEGAnnotator starts with a list dialog for the pipeline selection. The first choice will define the typology of analysis you want to perform.

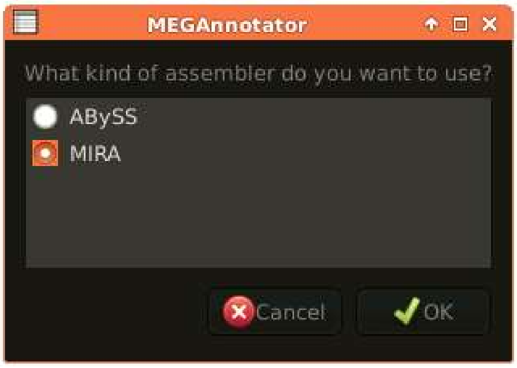


What kind of analysis do you want to perform?

○ Genomic_Assembly
○ Metagenomic_Assembly
○ Prediction_and_Annotation_only

a. Genomic Assembly

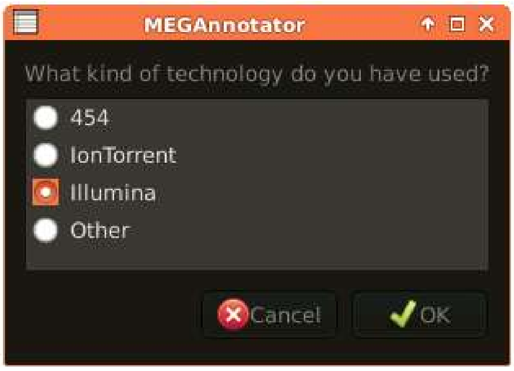
The genomic pipeline starts asking which assembler to use. In this tutorial, we show an example of genomic assembly using the MIRA assembler.



What kind of assembler do you want to use?

○ ABySS
● MIRA

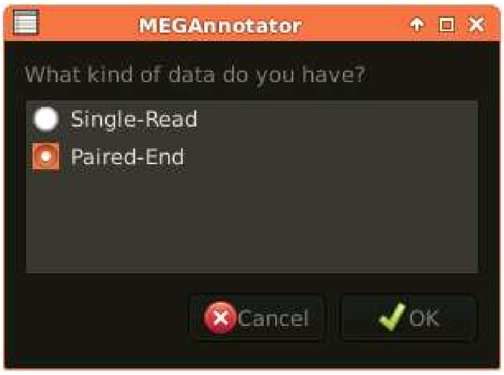
Then, MEGAnnotator needs the definition and the consequent selection of the raw reads that will be used from the assembler as input. In case the input is represented by Illumina data, a second list dialog allows the user to select paired- or single-end sequenced reads.



MEGAnnotator

What kind of technology do you have used?

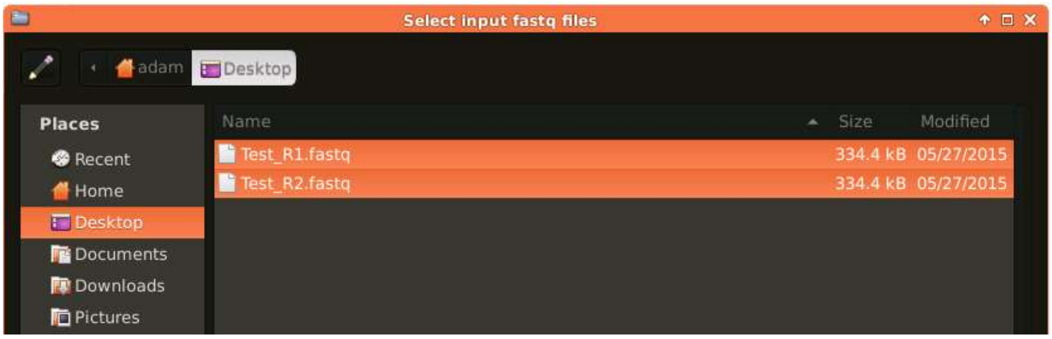
○ 454
○ IonTorrent
● Illumina
○ Other



MEGAnnotator

What kind of data do you have?

○ Single-Read
● Paired-End

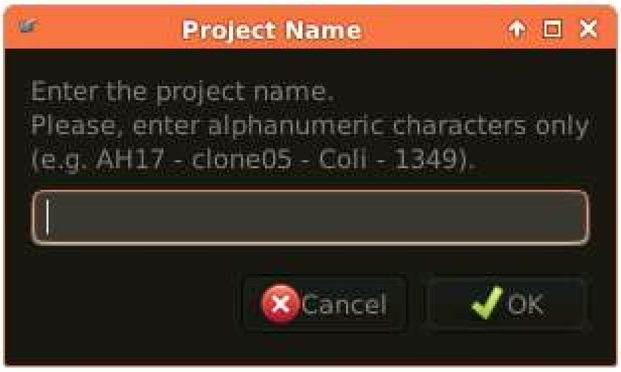


Select input fastq files

Name	Size	Modified
Test_R1.fastq	334.4 kB	05/27/2015
Test_R2.fastq	334.4 kB	05/27/2015

It is essential to provide the raw reads in fastq files, otherwise MEGAnnotator cannot manage the input. In the example, two illumina paired-end fastq files were selected.

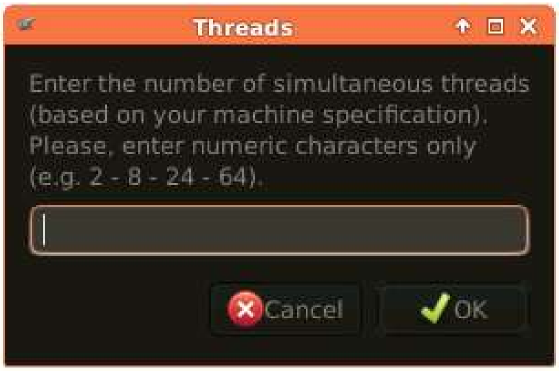
Consequentially, a text entry dialog awaits the project name to be insert. Please, enter alphanumeric characters only (e.g. AH17 – clone05 – Coli – 1349).



Project Name

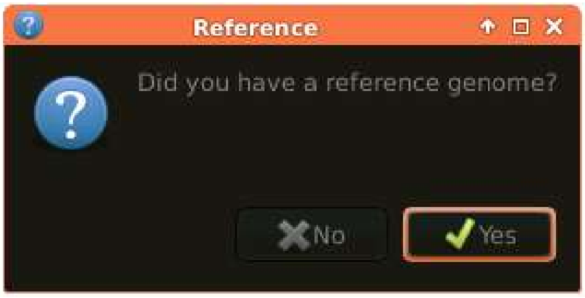
Enter the project name.
Please, enter alphanumeric characters only
(e.g. AH17 - clone05 - Coli - 1349).

Then, MEGAnnotator needs the number of threads you want allocate for the analyses. Please, enter numeric characters only (e.g. 2 – 8 – 24 – 64).

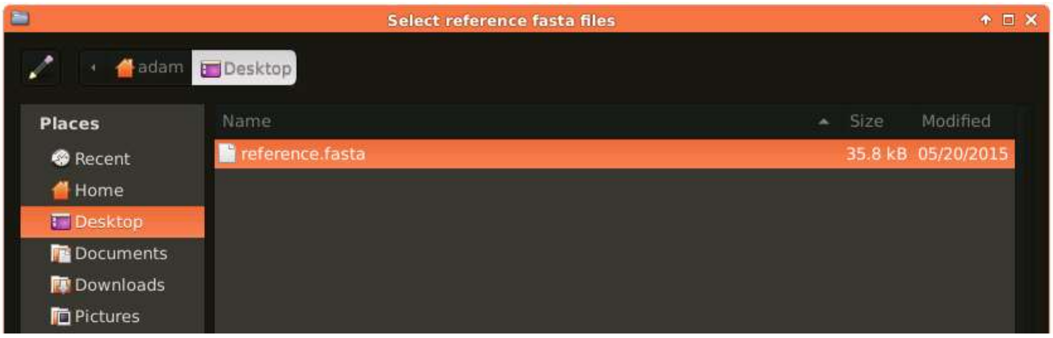


Enter the number of simultaneous threads (based on your machine specification). Please, enter numeric characters only (e.g. 2 - 8 - 24 - 64).

Following, you can give as input the genome reference you want to use for the contigs reordering after the assembly (optional). It is essential to provide the nucleotide sequence of the reference genome in fasta file.

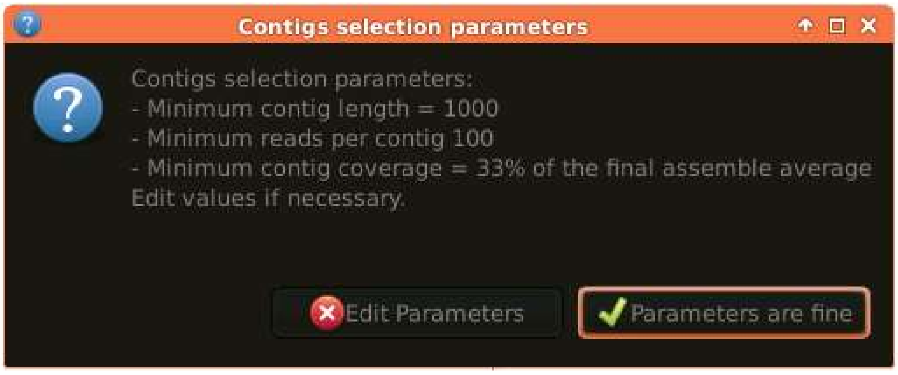


Did you have a reference genome?



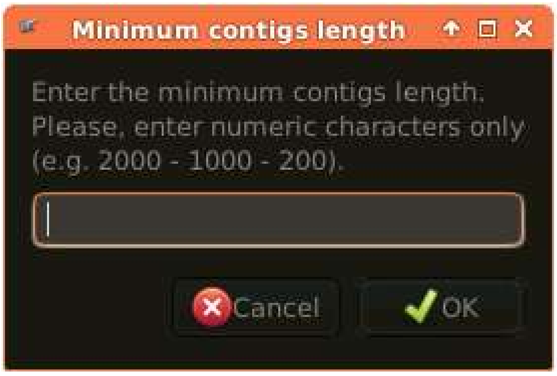
Name	Size	Modified
reference.fasta	35.8 kB	05/20/2015

Thereafter, the users can change the parameters set up for the contigs selection or continue with the parameter listed below.

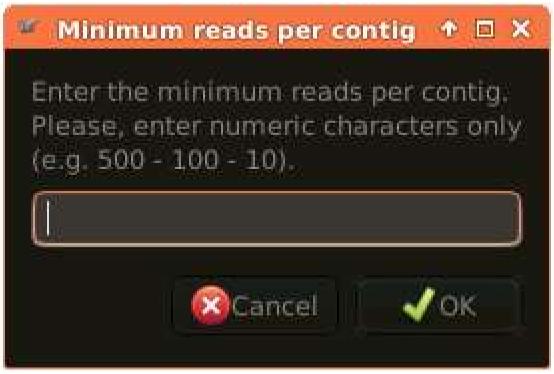


Contigs selection parameters:
- Minimum contig length = 1000
- Minimum reads per contig 100
- Minimum contig coverage = 33% of the final assemble average
Edit values if necessary.

Whether the user chose to edit the parameters, two text entry dialog gives the opportunity to set the minimum contig length and reads per contig. Please, enter numeric characters only.

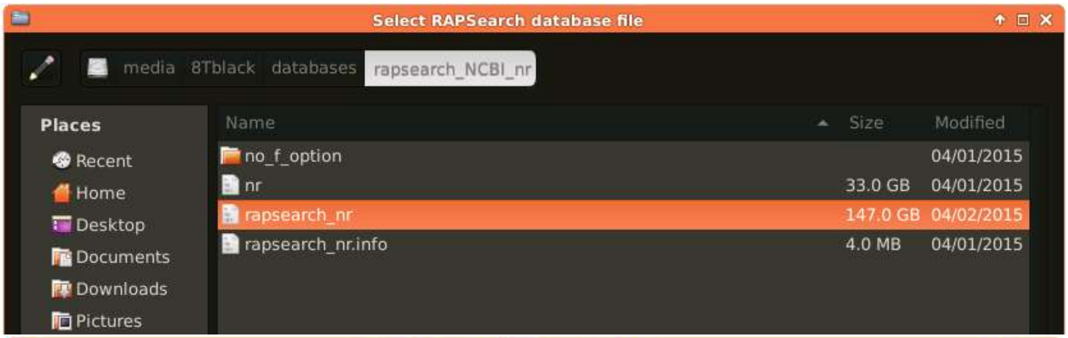


Enter the minimum contigs length.
Please, enter numeric characters only
(e.g. 2000 - 1000 - 200).

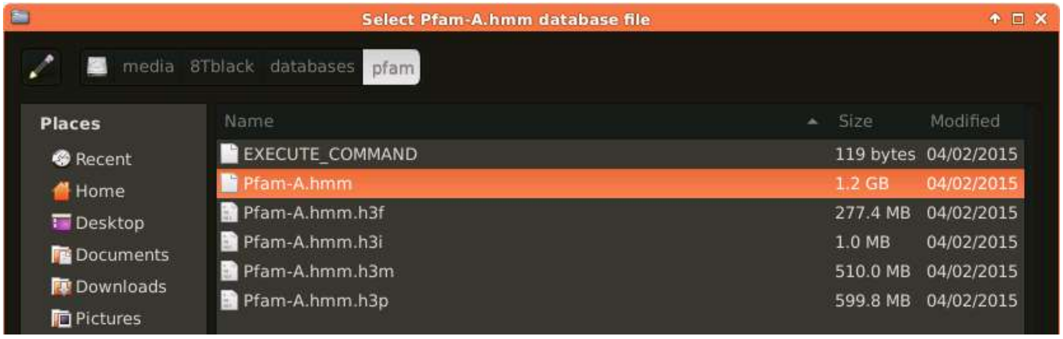


Enter the minimum reads per contig.
Please, enter numeric characters only
(e.g. 500 - 100 - 10).

In the two following steps, the user have to select the databases needed for the genome annotation. It is important to have already build the databases (see chapter 6).

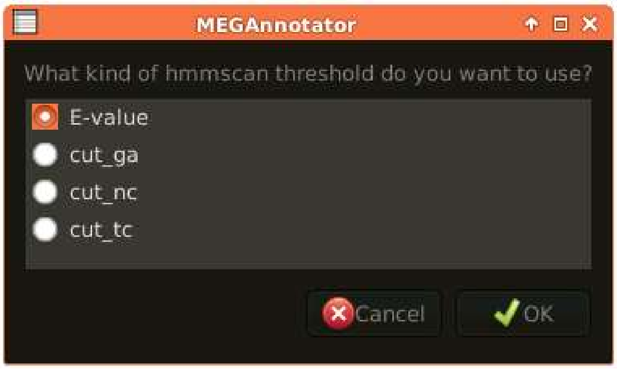


Name	Size	Modified
no_f_option		04/01/2015
nr	33.0 GB	04/01/2015
rapsearch_nr	147.0 GB	04/02/2015
rapsearch_nr.info	4.0 MB	04/01/2015



Name	Size	Modified
EXECUTE_COMMAND	119 bytes	04/02/2015
Pfam-A.hmm	1.2 GB	04/02/2015
Pfam-A.hmm.h3f	277.4 MB	04/02/2015
Pfam-A.hmm.h3i	1.0 MB	04/02/2015
Pfam-A.hmm.h3m	510.0 MB	04/02/2015
Pfam-A.hmm.h3p	599.8 MB	04/02/2015

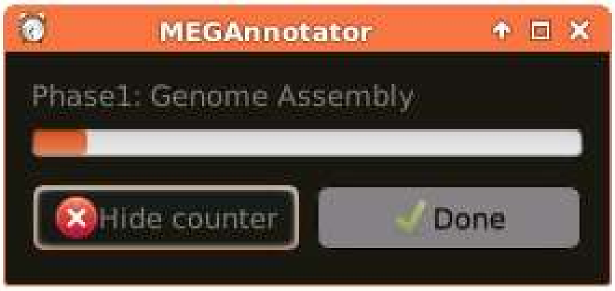
Then, MEGAnnotator needs to know if you want to choose a model-specific thresholding profile to apply in the HMM search (e-value cut-off of $1\text{x}10^{-10}$ was chosen as default).



What kind of hmmscan threshold do you want to use?

- ● E-value
- ○ cut_ga
- ○ cut_nc
- ○ cut_tc

After that, a progress dialog shows the progress status of the analysis.

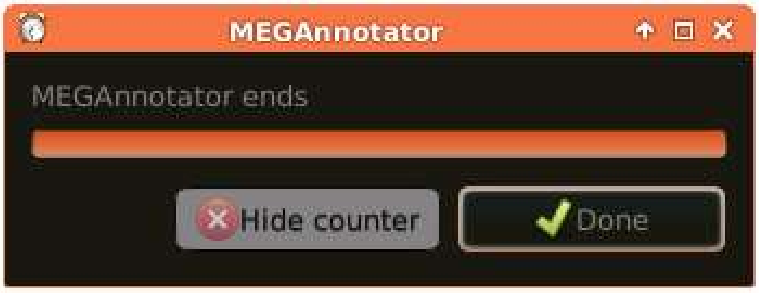


Phase1: Genome Assembly

Here, the list of the ten phases:

- Phase 1: Genome Assembly
- Phase 2: Contigs selection
- Phase 3: Alignment vs. reference genome
- Phase 4: Improvement of quality output
- Phase 5: Genes prediction
- Phase 6: RapSearch annotation
- Phase 7: pfam prediction
- Phase 8: merging annotation
- Phase 9: gbk generation and rRNA prediction
- Phase 10: genbank finalization and tRNA prediction

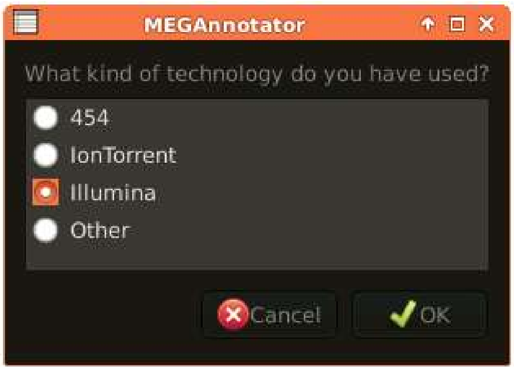
At the end, when all the phases ends correctly, the “Done” button will be available.



MEGAnnotator ends

b. Metagenomic Assembly

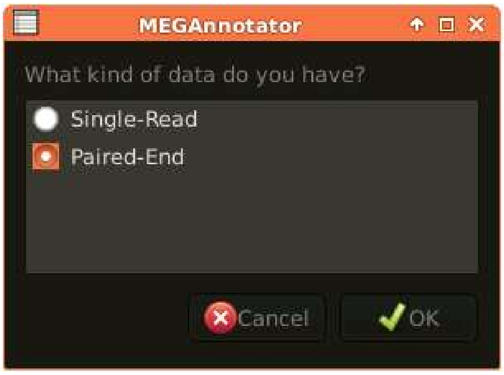
The metagenomic pipeline starts with the definition and selection of the raw reads that will be used from the assembler as input as well as the genomic pipeline presented above. In case the input is represented by Illumina data, a second list dialog allows the user to select paired or single end sequenced reads.



MEGAnnotator

What kind of technology do you have used?

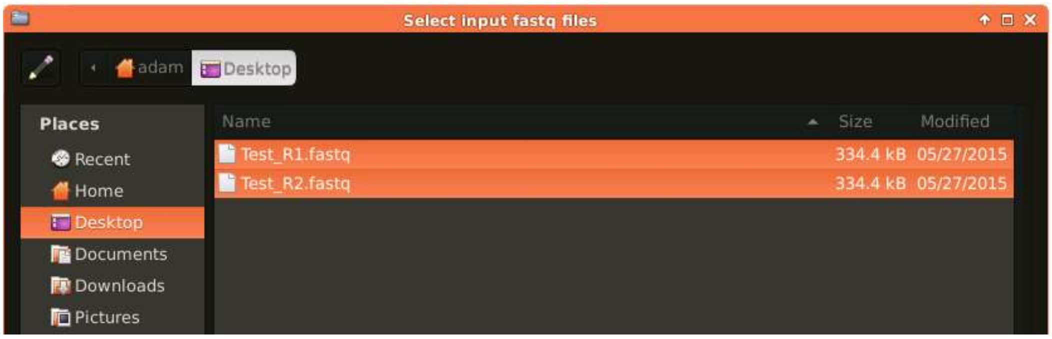
○ 454
○ IonTorrent
● Illumina
○ Other



MEGAnnotator

What kind of data do you have?

○ Single-Read
● Paired-End



Select input fastq files

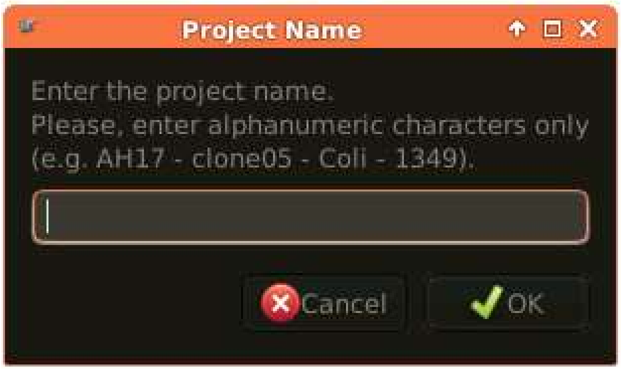
adam Desktop

Places
Recent
Home
Desktop
Documents
Downloads
Pictures

Name	Size	Modified
Test_R1.fastq	334.4 kB	05/27/2015
Test_R2.fastq	334.4 kB	05/27/2015

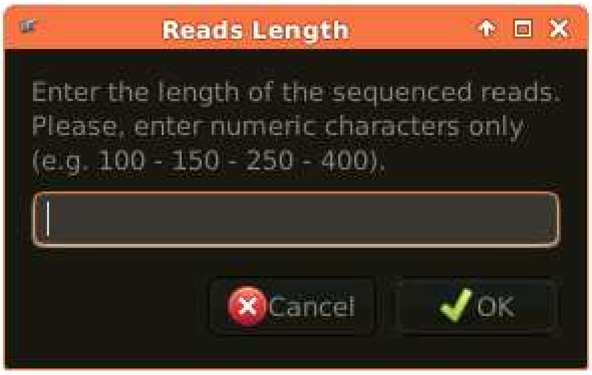
It is essential to provide the raw reads in fastq files, otherwise MEGAnnotator cannot manage the input. In the example, two illumina paired-end fastq files were selected.

Consequentially, two text entry dialog awaits the project name and the reads length to be insert. Please, enter alphanumeric characters only for the project name (e.g. AH17 – clone05 – Coli – 1349) and numeric characters only for the reads length (e.g. 100 – 150 – 250 – 400).



Project Name

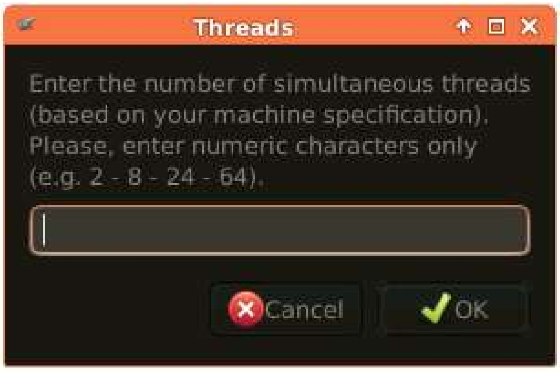
Enter the project name.
Please, enter alphanumeric characters only
(e.g. AH17 - clone05 - Coli - 1349).



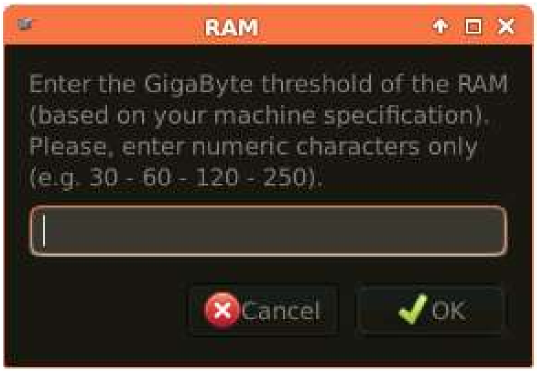
Reads Length

Enter the length of the sequenced reads.
Please, enter numeric characters only
(e.g. 100 - 150 - 250 - 400).

Then, MEGAnnotator needs the number of threads you want allocate for the analyses and the number of GigaByte of RAM the user wants to allocate. Please, enter numeric characters only.

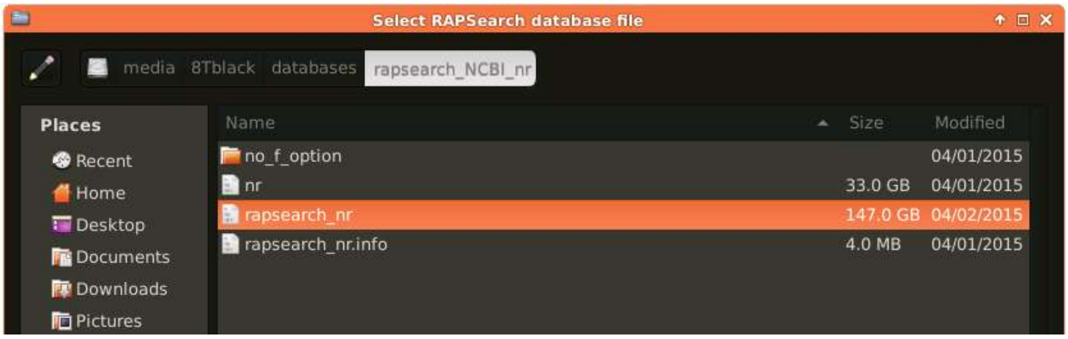


Enter the number of simultaneous threads (based on your machine specification). Please, enter numeric characters only (e.g. 2 - 8 - 24 - 64).



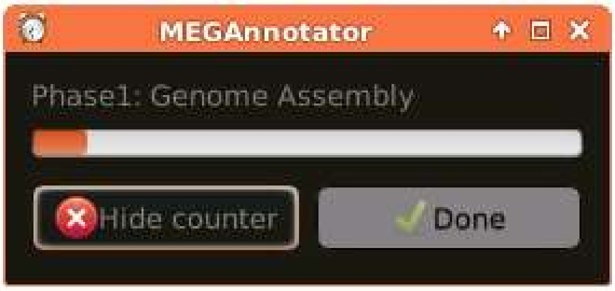
Enter the GigaByte threshold of the RAM (based on your machine specification). Please, enter numeric characters only (e.g. 30 - 60 - 120 - 250).

In the following step, the user have to select the databases needed for the genome annotation. It is important to have already build the NCBI nr database (see chapter 6).



Name	Size	Modified
no_f_option		04/01/2015
nr	33.0 GB	04/01/2015
rapsearch_nr	147.0 GB	04/02/2015
rapsearch_nr.info	4.0 MB	04/01/2015

After that, a progress dialog shows the progress status of the analysis.

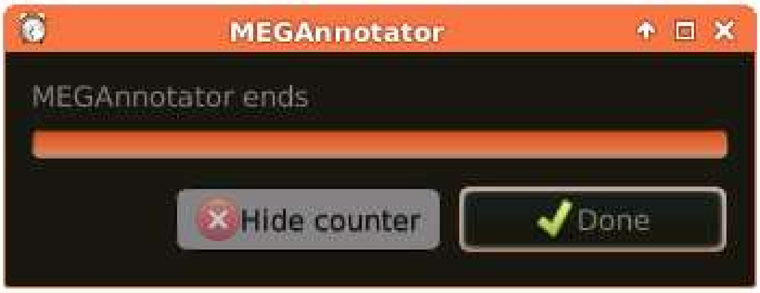


Phase1: Genome Assembly

Here, the list of the six phases:

- Phase 1: Metagenome Assembly
- Phase 2: Genes prediction
- Phase 3: RapSearch annotation
- Phase 4: merging annotation
- Phase 5: gbk generation and rRNA prediction
- Phase 6: genbank finalization and tRNA prediction

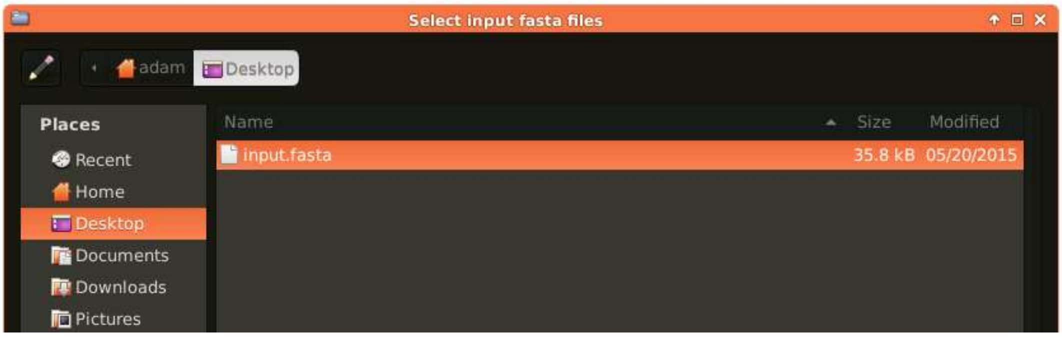
In the end, where all the phases ends correctly, the “Done” button will be available.



MEGAnnotator ends

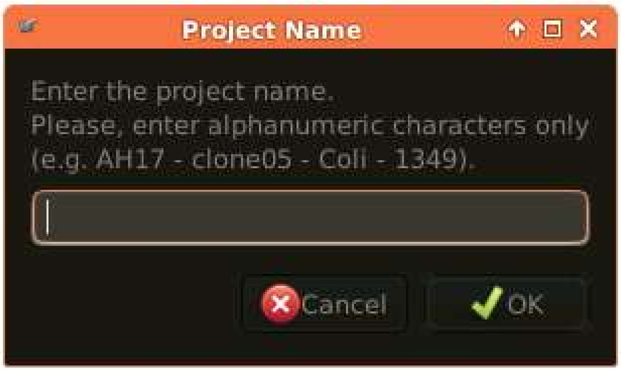
c. Genes Annotation only

The genes annotation pipeline starts with the selection of the multifasta file that the user wants to use as input.



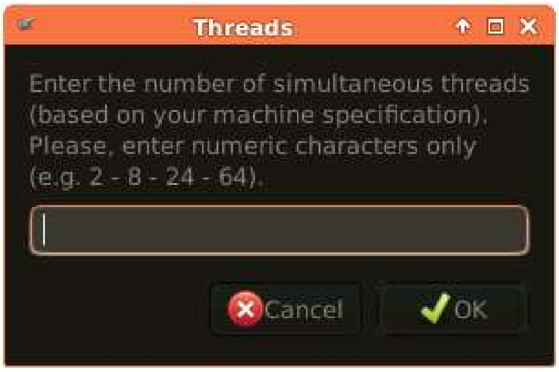
Name	Size	Modified
input.fasta	35.8 kB	05/20/2015

Consequentially, a text entry dialog awaits the project name to be inserted. Please, enter alphanumeric characters only (e.g. AH17 – clone05 – Coli – 1349).



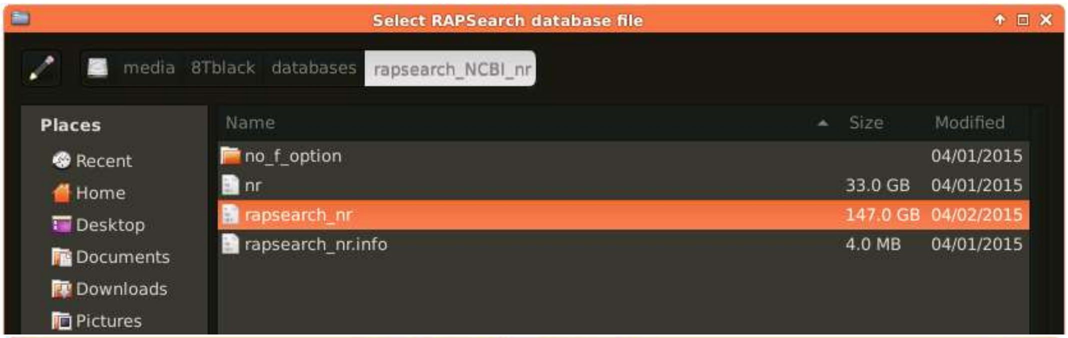
Enter the project name.
Please, enter alphanumeric characters only (e.g. AH17 - clone05 - Coli - 1349).

Then, MEGAnnotator needs the number of threads you want allocate for the analyses. Please, enter numeric characters only (e.g. 2 – 8 – 24 – 64).

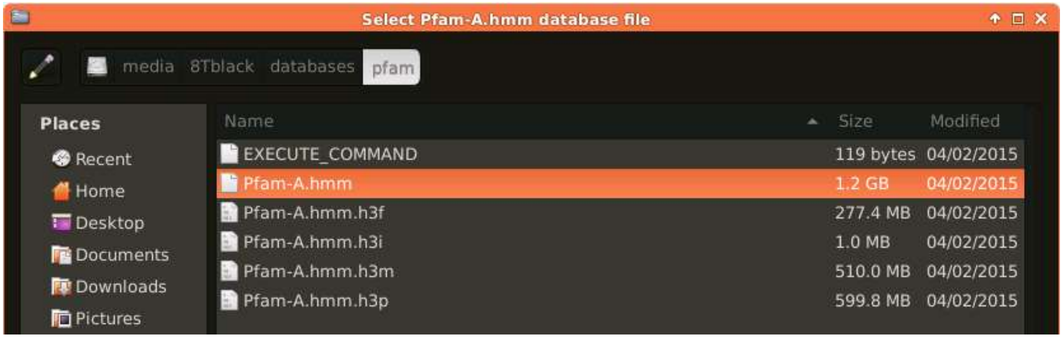

Threads

Enter the number of simultaneous threads (based on your machine specification). Please, enter numeric characters only (e.g. 2 - 8 - 24 - 64).

In the two following steps, the user have to select the databases needed for the genome annotation. It is important to have already build the databases (see chapter 6).

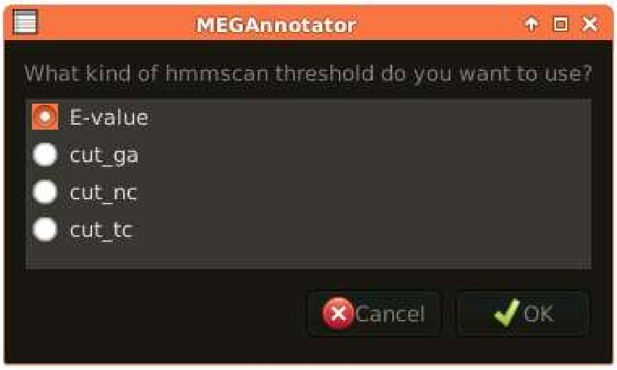

Select RAPSearch database file

Name	Size	Modified
no_f_option		04/01/2015
nr	33.0 GB	04/01/2015
rapsearch_nr	147.0 GB	04/02/2015
rapsearch_nr.info	4.0 MB	04/01/2015


Select Pfam-A.hmm database file

Name	Size	Modified
EXECUTE_COMMAND	119 bytes	04/02/2015
Pfam-A.hmm	1.2 GB	04/02/2015
Pfam-A.hmm.h3f	277.4 MB	04/02/2015
Pfam-A.hmm.h3i	1.0 MB	04/02/2015
Pfam-A.hmm.h3m	510.0 MB	04/02/2015
Pfam-A.hmm.h3p	599.8 MB	04/02/2015

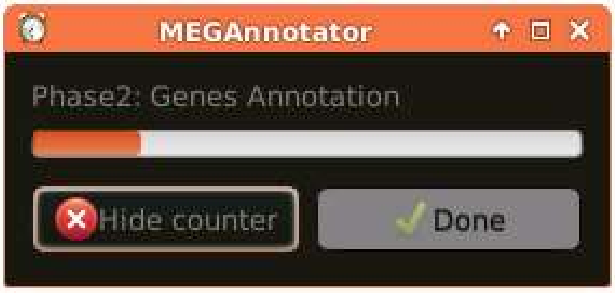
Then, MEGAnnotator needs to know if you want to choose a model-specific thresholding profile to apply in the HMM search (e-value cut-off of 1x10^{-10} was chosen as default).


MEGAnnotator

What kind of hmmscan threshold do you want to use?

● E-value
○ cut_ga
○ cut_nc
○ cut_tc

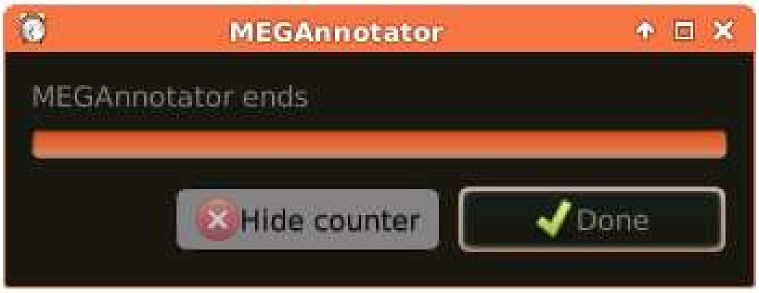
After that, the script starts showing a progress dialog.



Here, the list of the six phases:

- Phase 1: Genes prediction
- Phase 2: RapSearch annotation
- Phase 3: pfam prediction
- Phase 4: merging annotation
- Phase 5: gbk generation and rRNA prediction
- Phase 6: genbank finalization and tRNA prediction

In the end, where all the phases ends correctly, the “Done” button will be available.


MEGAnnotator
MEGAnnotator ends
Hide counter
Done